UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Editing

7-9 October 2024, Vienna

Enhancing Official Statistics through Artificial Intelligence: A Comparative Study of Imputation Techniques

Simona Cafieri, Mauro Bruno, Francesco Pugliese, Francesco Ortame,

Istat – Italian National Institute of Statistics, Italy

cafieri@istat.it, bruno@istat.it, frpuglie@istat.it francesco.ortame@istat.it

Abstract

In an increasingly globalized world, it is crucial to address emerging challenges related to health, environmental sustainability, and social inequalities with determination. These issues are deeply interconnected and require a comprehensive approach that actively involves National Statistical Institutes. These institutes are increasingly called upon to develop statistical frameworks that support informed policy decision-making. However, incomplete or missing data in questionnaires or registers can undermine the accuracy and reliability of the results.

The primary objective of this study is to evaluate the effectiveness of various imputation methods that leverage Machine Learning (ML) and Artificial Intelligence (AI) techniques to handle missing data in social surveys. To achieve this, a comparative analysis has been conducted, examining a range of imputation techniques from traditional statistical methods to advanced deep learning algorithms. These methods include Linear Regression (LR), k-nearest Neighbors (KNN), Decision Trees (DT), Random Forests (RF), Gradient Boosting (GB), Support Vector Machines (SVMs), and deep learning models such as Multi-Layer Perceptrons (MLPs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory networks (LSTMs).

All these methods are implemented as regressors to allow for an investigation of the full spectrum of regression-based imputation frameworks. The comparisons are based on real datasets from Istat's multipurpose household survey, where missing data are a common occurrence.

Preliminary results suggest that ML/AI-based imputation methods outperform traditional statistical techniques in terms of both performance and robustness, particularly when dealing with complex datasets and high-dimensional features.

This work, therefore, aims to explore innovative AI solutions to advance imputation techniques in official statistics, leading to more complete and accurate data on health, the environment, inequality, and other key areas.

1. Introduction

The advent of Artificial Intelligence (AI) is having a profound impact on a huge number of fields. In particular, in Official Statistics, the massive increase in non-traditional data sources, e.g., social network data, satellite images, and Internet of Things (IoT)-based sensor devices, raises real questions on the capacity of National Statistical Offices (NSOs) to use big data sources in statistical production. Some authors prove that Artificial intelligence (AI) can be used to overcome the issues related to automating data processing, improving data privacy and security, and enhancing the capabilities of IT human resources (Abbas, et. Al., 2023).

Furthermore, integrating AI methodologies offers innovative solutions to address data incompleteness, facilitating informed decision-making processes (Sun, et. Al., 2023). In recent years, National Statistical Institutes developed statistical frameworks across a range of domains to facilitate well-informed policy decisions. This is particularly pertinent in the context of today's increasingly interconnected world, where emerging challenges in environmental sustainability, health, and social inequalities require urgent attention (Rigo, 2022).

In Italy, these issues are of particular significance as they form the basis of the BES (equitable and sustainable well-being) indicators, which underpin the government's economic and financial planning document (Istat, 2024). It is worthy of note that a significant proportion of these indicators are based on survey data, which is susceptible to inaccuracy and unreliability if incomplete or missing responses are not addressed. The application of artificial intelligence, particularly machine learning (ML) and deep learning (DL) represents a promising solution to the issue of missing data in surveys. Such methods can be employed to predict and impute missing values, thereby enhancing the overall quality of statistical datasets. Conventional techniques, such as mean or median imputation, frequently introduce bias, whereas AI-based methodologies can facilitate more precise and impartial estimations. Several Machine Learning and Deep Learning models can be utilized for imputing missing data in health and environmental statistics.

2. Related works

The history of artificial intelligence is replete with the development of numerous algorithms, including support vector machines (SVM), which is a supervised learning algorithm that identifies the optimal hyperplane for the separation of different classes in the feature space. It is an effective algorithm for high-dimensional spaces and is commonly used for classification and regression (Suthaharan & Suthaharan, 2016). Some researchers have employed SVM to address the issue of missing values (Honghai et al., 2005).

Furthermore, we employed a decision tree, a non-linear model that divides data into branches based on feature values, to facilitate decision-making. It is beneficial for its interpretability and it is frequently employed in classification and regression tasks (Rokach, 2005). In their research, Nikfalazar (2020) employed both decision trees and fuzzy logic clustering for data imputation. Another noteworthy algorithm is XGBoost (Extreme Gradient Boosting), an advanced boosting algorithm that combines the predictions of multiple basic estimators to enhance model performance (Mitchell, 2017). It is renowned for its high efficiency and accuracy and is frequently employed in the resolution of numerous imputation problems (Rusdah & Murfi, 2020). Furthermore, we employed the k-nearest Neighbours (KNN) algorithm, which is a straightforward instance-based learning algorithm that classifies data points based on the majority class among their nearest neighbours. KNN is suitable for classification tasks in pattern recognition and recommendation systems (Guo et al., 2003). Furthermore, KNN has also been a popular choice for missing value imputation (Pujianto et al., 2019). Linear regression represents a methodology for modelling the relationship between a dependent variable and one or more independent variables. It is frequently employed for predictive analysis and trend forecasting (Montgomery et al., 2021). Random Forest (RF) is an ensemble learning method that constructs multiple decision trees and combines their outputs to enhance the accuracy and robustness of the resulting model. It is utilised in applications such as feature selection and complex classification tasks (Belgiu et al., 2016). Some authors have employed RF to impute missing values (Tang & Ishwaran, 2017). A Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that has been designed with the specific purpose of capturing long-term dependencies in sequential data, utilising memory cells. It is frequently employed for tasks such as speech recognition and time-series forecasting (Hochreiter & Schmidhuber, 1997). LSTM is also employed in the field of missing value imputation (Yuan et al., 2018). Finally, the Gated Recurrent Unit (GRU) is a type of Recurrent Neural Network (RNN) that is similar to the Long Short-Term Memory (LSTM) model but with a simplified architecture, resulting in increased speed and efficiency. GRU is applied in a variety of contexts, including natural language processing, sequential data analysis (Dey & Salem, 2017) and imputation (Wang et al., 2022).

This document aims to demonstrate how artificial intelligence (AI) can be employed to improve the quality of official statistics. In particular, the document examines the potential of machine learning (ML) and deep learning (DL) to enhance the accuracy, reliability, and comprehensiveness of health and environmental data.

3. Methods

One of the main sources of social and household health data in Italy is the Aspects of Daily Life (AVQ) survey, carried out annually by ISTAT.(ISTAT, 2022). AVQ represents an integral component of a unified system of social surveys. Indeed, collecting data is indispensable for understanding the daily lives of individuals and households. The survey provides information on the habits of citizens and the problems they face in everyday life through interviews with a sample of 20,000 households, representing approximately 50,000 individuals. Since 2018, the survey has been executed using a sequential CAWI/PAPI mixed-mode technique.. The survey investigates a range of social aspects, including education, employment, family and social life, leisure time, political and social participation, health, lifestyles, access to services and other factors relevant to the study of quality of life.. These topics are investigated from a social perspective, with particular consideration given to behaviours, motivations and opinions as key elements in the definition of social information. The survey is included in the National Statistics Plan, which collates the statistical investigations that are required for the country. However, it is not uncommon for questionnaires to be incomplete, which can affect the precision and dependability of the resulting data.

To address this issue, we have designed and implemented an imputation experiment by using the AVQ dataset from the 2021 survey comprising 735 variables. The presence of missing values in the dataset, frequently represented as blanks or NaN, is incompatible with scikit-learn estimators, which require all values to be numeric and significant. A fundamental approach is complete case analysis, whereby rows (dropNA) or columns with missing values (list-wise deletion) are excluded. Nevertheless, this may result in a significant reduction of the available information. An effective strategy is to impute missing data by inferring it from the available data. Conversely, techniques such as the use of central tendency measures (mean, median, etc.) can be employed. This approach appears to be relatively straightforward and robust. However, there is a risk of underestimating or overestimating the true values, which could introduce bias into the resulting estimates. This phenomenon occurs when an algorithm produces results that are systematically biased due to incorrect assumptions, which are typically present in the data set or in the machine learning process.

In this study, we use missing data imputation techniques known as 'regression imputation'. Essentially, this method estimates missing values using a regressor (e.g. support vector regressor or random forest regressor), with the missing variable as the target and the other variables as inputs. Regression imputation is divided into 'deterministic' and 'stochastic'. The main difference between these two approaches is how the missing values are estimated and how uncertainty is taken into account. In deterministic regression imputation, missing values are estimated using a deterministic relationship between the variables. The trained regression model is used to predict the missing values for incomplete observations. The predicted value is used directly as an estimate for the missing value, hence the term deterministic. Deterministic imputation does not take into account the uncertainty associated with the estimate, so the predicted values are always the same for a given combination of input values. In contrast, stochastic regression imputation incorporates the uncertainty of the estimate into the imputation process. A stochastic noise term is added to the prediction. This noise term can be generated using the distribution of residual errors from the regression model. For example, if the regression model has a residual variance of sigma squared, noise can be added extracting it from a normal distribution with a mean of 0 and a variance of sigma squared, resulting in a more realistic estimate. However, in our work we only used deterministic regression imputation, with the intention of exploring stochastic imputation adapted to machine learning and deep learning models in the future. As a final observation, we cannot directly impute regression values before preprocessing. The input predictor variables also contain missing data themselves, which would cause issues for the machine learning and deep learning models, as the libraries we used (i.e., Scikit-Learn or Keras) do not accept null values. Therefore, we imputed all input variables with missing values using a method called "Simple Random Imputation," which involves replacing the missing value with a random value. It is proven that this approach does not significantly affect the final estimate given the large number of variables present in our dataset.

In the literature, this method is more efficient than other methods of replacing predictor variables with a zero, the mean, etc. (Kalton & Kish, 1984).

In this work, we trained all the traditional machine learning models described in the previous section: SVM, DT, RF, XGBoost, KNN, and the most recent deep learning models: MLP, LSTM, GRU, CONV1D. The objective of the training was to create models for the imputation of the following health-related variables: The variables of interest were body mass index (BMI) for individuals aged 18 and over. The same models were trained for the imputation of additional environmental variables, namely SODPOAP (**resident satisfaction with household**

waste collection services). The ML and DL models have been evaluated using a range of metrics that are appropriate for regression problems. The Root Mean Squared Error (RMSE), the Mean Absolute Percentage Error (MAPE), the Mean Absolute Error (MAE) and the R2 Score were employed for the assessment of the models. It was observed that even though the RMSE yields absolute values, it was an adequate metric for comparing the performance of the different models and for their ranking. Accordingly, the trained models were ordered in descending order of RMSE, with the most effective model identified as the one with the lowest error. Furthermore, models that demonstrated minimal overfitting, as evidenced by a minimal discrepancy between the RMSE values for the training and test sets, were deemed the most optimal.

4. Results

Upon completion of the training phase, it was observed that the combination of the most effective models varied depending on the variable being imputed. This finding may be related to the No-Free-Lunch Theorem, which sets a theoretical limit in machine learning. The No-Free-Lunch Theorem postulates that no single optimal machine learning model exists for every task. Consequently, the strength of our method lies in its capacity to identify the optimal model for each variable (for each variable, the task is completely distinct) to be imputed, which is meticulously selected from a vast array of models. About the health-related variable "BMI", the results for all the models are presented in Table 1. It can be seen that the Deep Learning model is the most effective. Long Short-Term Memory (LSTM) model. Table 2 presents a comparison of the descriptive statistics (mean, standard deviation, and quartiles) calculated for the original variable before imputation and the imputed variable. This preliminary assessment indicates that the imputed distribution is not markedly disparate from the original distribution.

Table 1 – *Table of metrics of the models trained to predict the variable "BMI".*

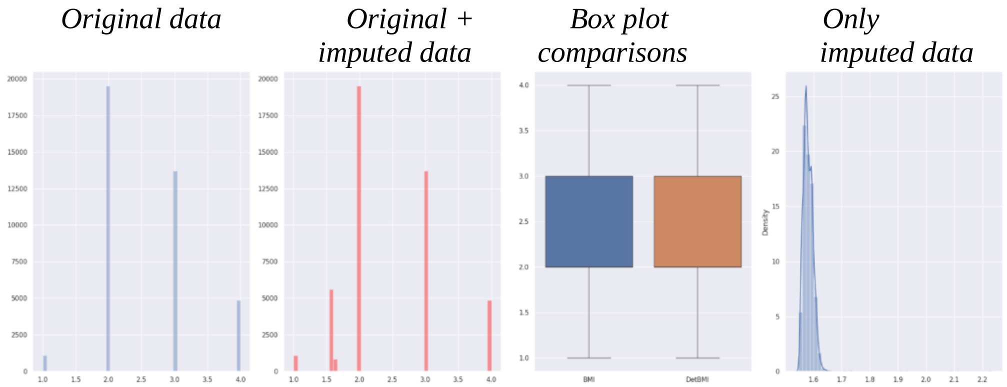
MODEL	Training RMSE	Test RMSE
LSTM	0.7489	0.7546
GRU	0.7423	0.7551
CONV1D	0.7637	0.7854
MLP	0.7866	0.8122
SVM	0.8524	0.8823
KNN	0.6954	0.7949
LR	0.6400	0.7936
XG Boost	0.3898	0.7318
RF	0.2666	0.7240
DT	0.0000	1.0143

Table 2 – *Table of comparisons between descriptive statistics of BMI and Imputed BMI (DetBMI).*

MODEL	Mean	STD	MIN	25%	50%	75%	MAX
BMI	2.5695	0.7413	1	2	2	3	4
DetBMI	2.4296	0.7683	1	2	2	3	4

Figure 1 provides a comparison between the original distributions (omitting the nulls in the BMI column) and the distribution with imputed data (the BMI column without nulls plus imputed values). Furthermore, an analysis of the box plots of both distributions and the distribution of only the imputed values that replace the nulls is provided. As can be observed, the two distributions are similar, although a slight margin of error is to be expected at this stage.

Figure 1 – *Comparisons among distributions charts for the variable BMI.*



Original data

Original + imputed data

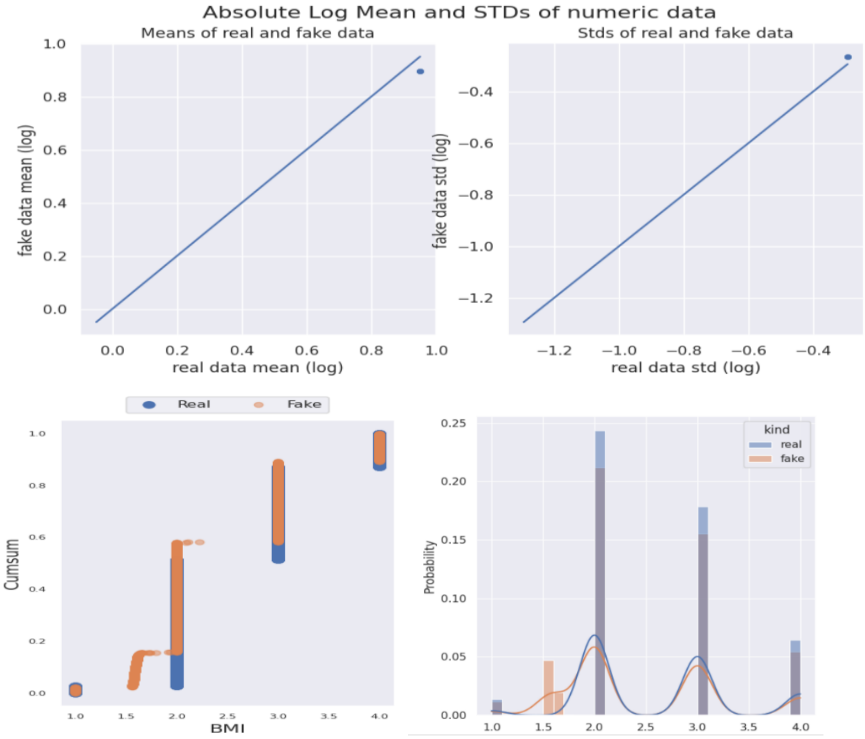
Box plot comparisons

Only imputed data

Figure 2 presents a further comparison between the mean and standard deviation c imputed data and the original data. Furthermore, a comparison between the original an puted univariate and cumulative distributions is presented. The distributions are ighly lar, which reinforces the assertion that AI models can markedly enhance the handlir missing data.

.

Figure 2 – *Comparisons among means, standard deviations, cumulate and univariate distributions charts*



Regarding the environmental variable SODPOAP, Table 3 provides a comparison of the metrics for all models, indicating that the Multi-Layer Perceptron (MLP) is the optimal model in this context. The most recent models exhibit a proclivity for overfitting, as evidenced by the markedly lower error rate on the training set in comparison to the test set. It is also noteworthy that, as anticipated, deep learning models demonstrate superior performance compared to machine learning models on these high-dimensional imputation datasets. Indeed, deep learning models consistently rank among the top performers, irrespective of whether they are recurrent or not. This suggests that the longitudinal (temporal) aspect of the data does not influence the models' performance in this dataset. In comparison to traditional machine learning models, simpler models such as SVM and LR appear to demonstrate superior performance, whereas more sophisticated ensemble models like XGBoost and RF tend to exhibit a higher propensity for overfitting

Table 3 – *Table of metrics of the models trained to predict the variable "SODPOAP".*

MODEL	Training RMSE	Test RMSE
MLP	0.6843	0.6970
GRU	0.6048	0.6190
CONV1D	0.6169	0.6313
LSTM	0.6020	0.6165
SVM	0.6031	0.6189
KNN	0.5347	0.6115
LR	0.5093	0.6195
XGBoost	0.2915	0.5315
RF	0.1964	0.5337
DT	0.0000	0.7211

Comparisons of descriptive statistics are shown in Table 4.

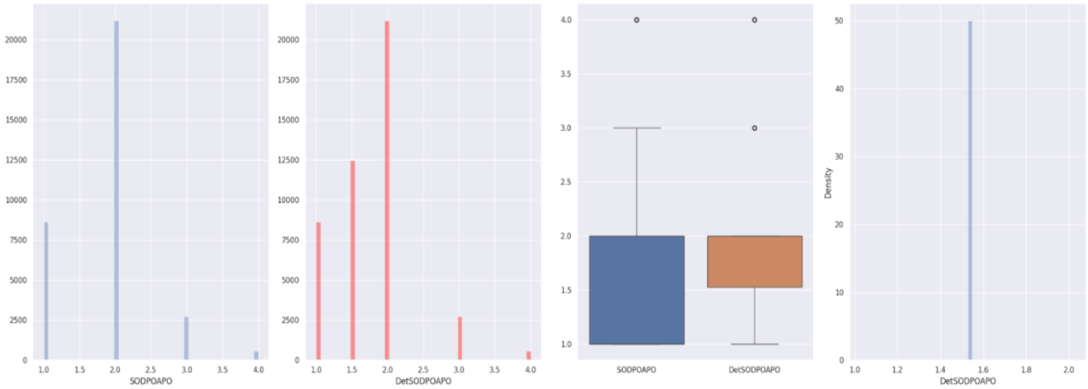
Table 4 – *Table of comparisons between descriptive statistics of SODPOAP and Imputed SODPOAP (DetSODPOAP).*

MODEL	Mean	STD	MIN	25%	50%	75%	MAX
SODPOAP	1.8563	0.6260	1	1	2	2	4
DetSODPOAP	1.7666	0.5532	1	1.5	2	2	4

Figure 3 presents a comparison of the original and imputed distributions of SODPOAP, while Figure 4 provides a comparison of the means, standard deviations, univariate, and cumulative distributions. As with the previous results, excellent results are obtained. However, it can be observed that the outcome and behaviour of the models change depending on the difficulty level of the variable. The principal advantage of our methodology is the construction of a bespoke model for each variable to be imputed.

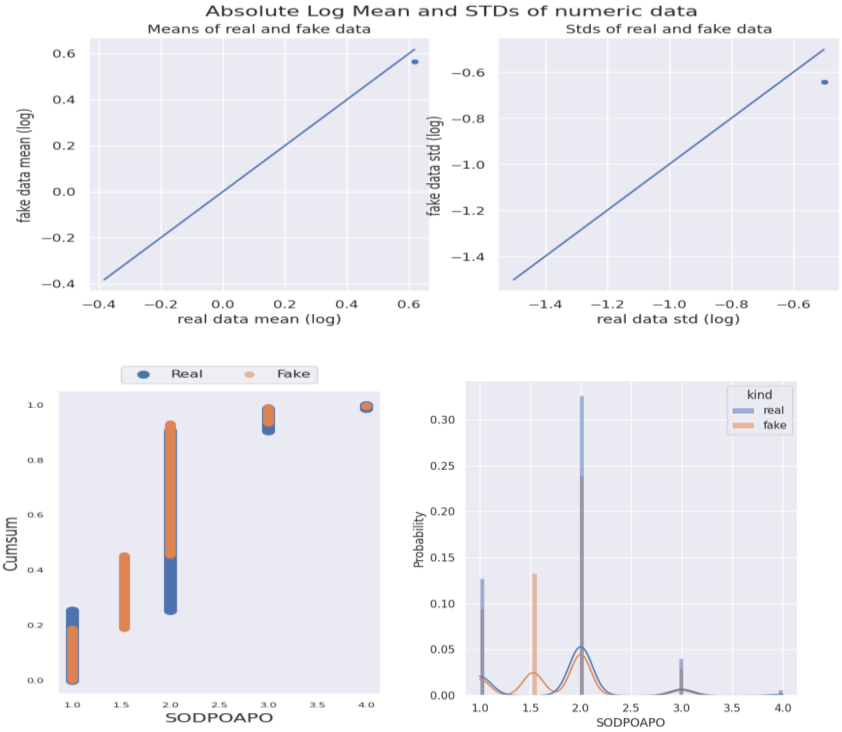
Figure 3 – *Comparisons among distributions charts for the variable SODPOAP*

. *Original data* *Original and imputed data* *Box plot comparisons* *Only imputed data*



The results demonstrate that deep learning models, such as long short-term memory (LSTM) and gated recurrent unit (GRU) networks, exhibit high performance in terms of root mean square error (RMSE) on every task, indicating their suitability for handling sequential data. Random Forest and XGBoost also demonstrated satisfactory performance, however, they tended to overfit, rendering them unsuitable for the task of the imputation of missing values. The support vector machine models demonstrated reliable performance in imputation, although they exhibited slightly higher root mean square error (RMSE) compared to the top-performing models.

Figure 4 *Comparisons among means, standard deviations, cumulate and univariate distributions charts*



5. Final remarks

The analysis of health and environmental statistics presents a promising avenue for enhancing the quality and reliability of data. Artificial intelligence (AI) —-based imputation methods, particularly those involving machine learning (ML) and deep learning (DL) techniques, can effectively address the issue of missing data, thereby enhancing the overall integrity of statistical surveys. A comprehensive analysis enables the selection of the optimal model, the imputation of missing data using this model, and the subsequent evaluation of the quality of the imputed data. Future research should prioritise the development of a stochastic regression imputation method that more effectively preserves the variance between the original and imputed distributions, and the investigation of advanced models, such as Transformers and Generative Adversarial Networks (GANs),to further enhance the imputation process

References

ABBAS, S. W., HAMID, M., ALKANHEL, R., & ABDALLAH, H. A. (2023). *Official Statistics and Big Data Processing with Artificial Intelligence: Capacity Indicators for Public Sector* Organizations. Systems, 11(8), 424.

BELGIU, M., & DRĂGUŢ, L. (2016). *Random forest in remote sensing: A review of applications and future directions*. ISPRS journal of Photogrammetry and remote sensing, 114, 24-31

DEY, R., & SALEM, F. M. (2017, August). Gate-variants of gated recurrent unit (GRU) neural networks. In *2017* IEEE 60th International Midwest Symposium on Circuits and systems (MWSCAS) (pp. 1597-1600). IEEE.

GUO, G., WANG, H., BELL, D., BI, Y., & GREER, K. (2003). *KNN model-based approach in classification*. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings (pp. 986-996). Springer Berlin Heidelberg.

HOCHREITER, S., & SCHMIDHUBER, J. (1997). *Long short-term memory. Neural Computation, 9*(8), 1735-1780.

ISTAT. (2022). Indagine Aspetti della vita quotidiana 2021.

ISTAT, (2024). Il Rapporto BES 2023

LUENGO J., GARCÍA S., HERRERA F. *On the choice of the best*

MARAVIGLIA, L. (2022), *INVALSI data: a source to improve our knowledge of digital divide among students. use of invalsi data in school.*

MITCHELL, R., & FRANK, E. (2017). *Accelerating the XGBoost algorithm using GPU computing.* PeerJ Computer Science, 3, e127.

MONTGOMERY, D. C., PECK, E. A., & VINING, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

NIKFALAZAR, S., YEH, C. H., BEDINGFIELD, S., & KHORSHIDI, H. A. (2020). *Missing data imputation using decision trees and fuzzy clustering with iterative learning*. Knowledge and Information Systems, 62, 2419-2437.

PUJIANTO, U., WIBAWA, A. P., & AKBAR, M. I. (2019, October). *K-nearest neighbor (k-NN) based missing data imputation.* In 2019 5th International Conference on Science in Information Technology (ICSITech) (pp. 83-88). IEEE.

RIGO, A. (2022). *Programmazione e innovazione: il percorso verso l'efficienza interna delle Pubbliche Amministrazioni*.

ROKACH, L., & MAIMON, O. (2005). *Decision trees. Data mining and knowledge discovery handbook*, 165-192

RUSDAH, D. A., & MURFI, H. (2020). *XGBoost in handling missing values for life insurance risk prediction*. SN Applied Sciences, 2(8), 1336.

SUN, Y., LI, J., XU, Y., ZHANG, T., & WANG, X. (2023). *Deep learning versus conventional methods for missing data imputation: A review and comparative study*. Expert Systems with Applications, 227, 120201.

SUTHAHARAN, S., & SUTHAHARAN, S. (2016). *Support vector machine. Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, 207-235.

TANG, F., & ISHWARAN, H. (2017). *Random forest missing data algorithms*. Statistical Analysis and Data Mining: The ASA Data Science Journal, 10(6), 363-377.

YUAN, H., XU, G., YAO, Z., JIA, J., & ZHANG, Y. (2018, October). *Imputation of missing data in time series for air pollutants using long short-term memory recurrent neural networks*. In Proceedings of the 2018 ACM international joint conference and 2018 international symposium on pervasive and Ubiquitous Computing and wearable computers (pp. 1293-1300).

WANG, R., ZHANG, Z., WANG, Q., & SUN, J. (2022). T*ime and location gated recurrent unit for multivariate time series imputation*. EURASIP Journal on Advances in Signal Processing, 2022(1), 74.

KALTON, G., & KISH, L. (1984). *Some efficient random imputation methods*. Communications in Statistics-Theory and Methods, *13*(16), 1919-1939.